MATH 523 A1 Q3

2024-02-13

Question 3, part (a)

```
library(palmerpenguins)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data(penguins)
penguins_complete<-penguins %>% drop_na()
penguins$sex <- ifelse(penguins$sex=="male", 0, 1)
model<-glm(sex ~ body_mass_g, data = penguins_complete, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = sex ~ body_mass_g, family = "binomial", data = penguins_complete)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7486  -0.9214   0.4292   1.1067   1.6806
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.1625416  0.7243906  -7.127 1.03e-12 ***
## body_mass_g  0.0012398  0.0001727   7.177 7.10e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 461.61  on 332  degrees of freedom
## Residual deviance: 396.64  on 331  degrees of freedom
## AIC: 400.64
##
## Number of Fisher Scoring iterations: 4
```

Question 3, part (b)

null: body_mass_g = 0

alternative: body_mass_g != 0

significance level: alpha = 0.05

Since the p-value (for the slope coefficient) is 7.10e-13 which is less than 0.05, we reject the null hypothesis. This means that there is a significant relationship between body mass of the birds and their sex (the slope of the coefficient is not zero, even though it is pretty close to zero).

Question 3, part (c)

The coefficient (beta1_hat) for body_mass_g is 0.0012398, therefore exp(beta1_hat) = e^(0.0012398) which is around 1.00124

Therefore a penguin's weighing one gram more than another is very slightly more likely to be categorized as male than the other penguin. The odds ratio is very close to 1, so a change in body mass has a small effect on the likelihood of being male versus female.

Question 3, part (d)

```
library(palmerpenguins)
library(tidyverse)
data(penguins)
penguins_complete<-penguins %>% drop_na()
penguins$sex <- ifelse(penguins$sex=="male", 0, 1)
model<-glm(sex ~ body_mass_g, data = penguins_complete, family = "binomial")
fitted_prob<-predict(model, newdata=data.frame(body_mass_g=3500), type="response")
fitted_se<-predict(model, newdata = data.frame(body_mass_g = 3500), type = "response", se.fit = TRUE)
ci <- fitted_se$fit + c(-1, 1) * qnorm(0.975) * fitted_se$se.fit

fitted_prob
```

```
##         1
## 0.3050902
```

```
ci
```

```
## [1] 0.2371929 0.3729876
```

The fitted probability that a penguin is male penguin if the penguins weight is 3500g is 0.3050902.

The 95% confidence interval is [0.2371929, 0.3729876]

Question 3, part (e)

```
library(palmerpenguins)
library(tidyverse)
data(penguins)
penguins_complete<-penguins %>% drop_na()
penguins$sex <- ifelse(penguins$sex=="male", 0, 1)
model<-glm(sex ~ body_mass_g, data = penguins_complete, family = "binomial")

library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

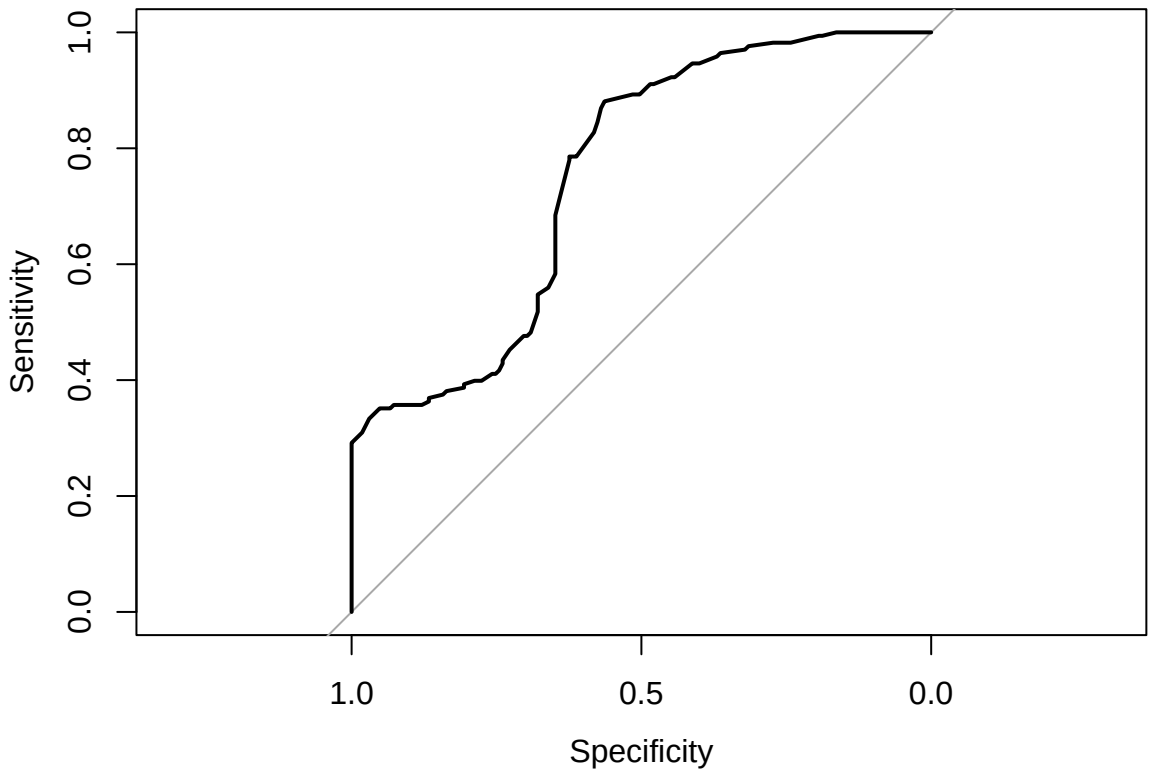
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
probabilities <- predict(model, type = "response")
roc_curve <- roc(penguins_complete$sex, probabilities)
```

```
## Setting levels: control = female, case = male
```

```
## Setting direction: controls < cases
```

```
plot(roc_curve)
```




```
auc(roc_curve)
```

```
## Area under the curve: 0.752
```

The model predicts the sex of a penguin based on its body mass with an accuracy of 75.2%, which is better than guessing (50% chance), but it is not close to perfect as there is a lot of room for error.