HW 11 Question 1 Solutions

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

Data<-penguins

##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]

##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
```

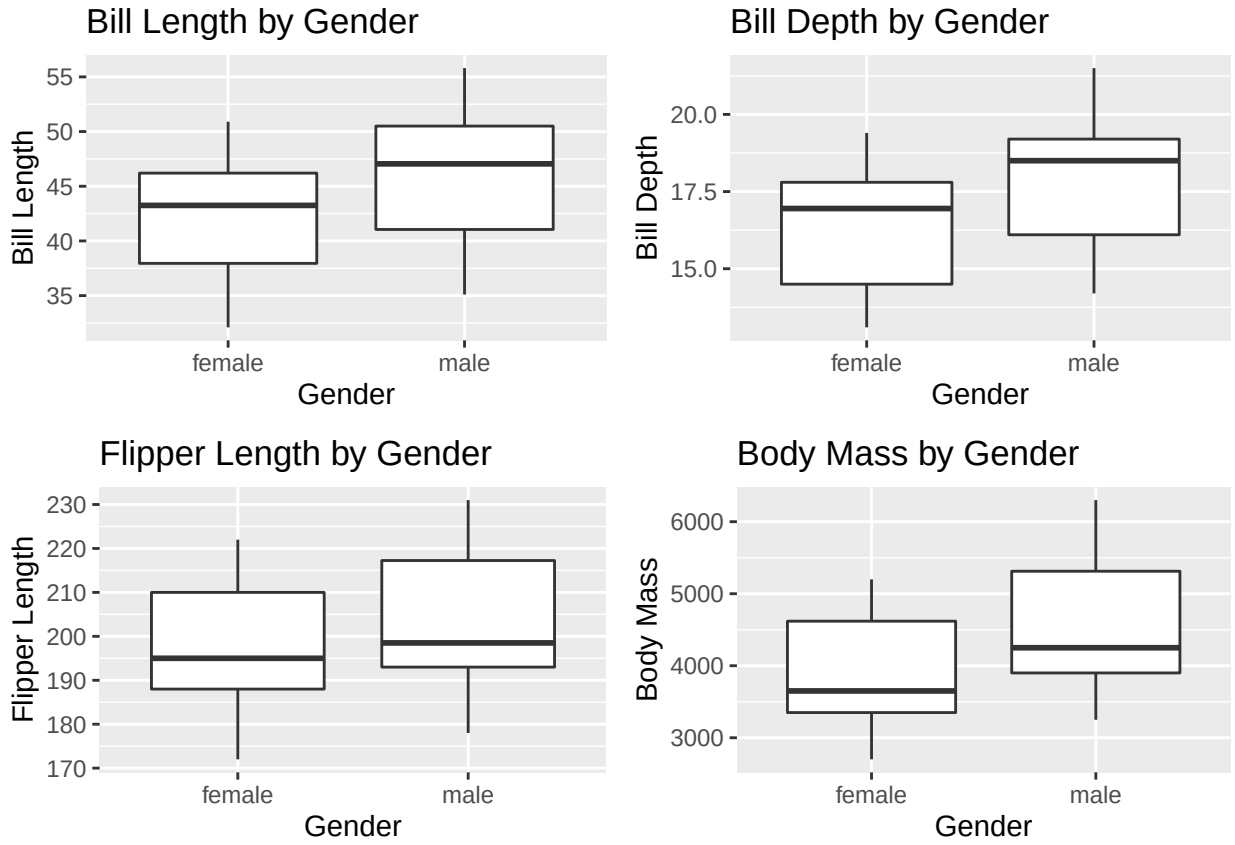
1)

(a)

We create boxplots to compare the distributions of the quantitative variables between genders.

```
bp1<-ggplot(train, aes(x=sex, y=bill_length_mm))+
  geom_boxplot()+
  labs(x="Gender", y="Bill Length", title="Bill Length by Gender")

bp2<-ggplot(train, aes(x=sex, y=bill_depth_mm))+
  geom_boxplot()+
  labs(x="Gender", y="Bill Depth", title="Bill Depth by Gender")

bp3<-ggplot(train, aes(x=sex, y=flipper_length_mm))+
  geom_boxplot()+
  labs(x="Gender", y="Flipper Length", title="Flipper Length by Gender")

bp4<-ggplot(train, aes(x=sex, y=body_mass_g))+
```

```
  geom_boxplot()+
  labs(x="Gender", y="Body Mass", title="Body Mass by Gender")

##produce the 4 boxplots in a 2 by 2 matrix
grid.arrange(bp1, bp2, bp3, bp4, ncol = 2, nrow = 2)
```



We can see that male penguins tend to have longer bill lengths, deeper bill depths, longer flippers, and larger body masses. Male penguins tend to be bigger than female penguins.

(b)

```
result<-glm(sex ~ ., family="binomial", data=train)
summary(result)
```

```
##
## Call:
## glm(formula = sex ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -2.85959  -0.10720   0.00061   0.06817   3.02072
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -94.355394  17.638204  -5.349 8.82e-08 ***
## speciesChinstrap  -10.608813   2.634752  -4.026 5.66e-05 ***
## speciesGentoo     -10.384568   3.565641  -2.912  0.00359 **
## bill_length_mm      1.025200   0.238593   4.297 1.73e-05 ***
## bill_depth_mm       2.287977   0.516595   4.429 9.47e-06 ***
## flipper_length_mm  -0.088318   0.065040  -1.358  0.17450
## body_mass_g         0.008094   0.001662   4.871 1.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 368.619  on 265  degrees of freedom
## Residual deviance:  68.297  on 259  degrees of freedom
## AIC: 82.297
##
## Number of Fisher Scoring iterations: 8
```

Flipper length is the only predictor with an insignificant Wald test. So we can drop it from our logistic regression.

(c)

```
reduced<-glm(sex~species+bill_length_mm+bill_depth_mm+body_mass_g,
             family="binomial", data=train)
reduced
```

```
##
## Call:  glm(formula = sex ~ species + bill_length_mm + bill_depth_mm +
##     body_mass_g, family = "binomial", data = train)
##
## Coefficients:
##      (Intercept)  speciesChinstrap     speciesGentoo    bill_length_mm
##       -1.032e+02        -1.042e+01        -1.238e+01         9.513e-01
##    bill_depth_mm       body_mass_g
##        2.099e+00         7.714e-03
##
```

```
## Degrees of Freedom: 265 Total (i.e. Null);  260 Residual
## Null Deviance:       368.6
## Residual Deviance: 70.17     AIC: 82.17
```

The logistic regression is

$$\log \frac{\hat{\pi}}{1-\hat{\pi}} = -103.2 - 10.42I_1 - 12.38I_2 + 0.9513x_1 + 2.099x_2 + 0.007714x_3$$

where $I_1 = 1$ for a Chinstrap penguin, 0 otherwise, $I_2 = 1$ for a Gentoo penguin, 0 otherwise, and x_1, x_2, x_3 denote bill length, bill depth, and body mass respectively.

(d)

Since the coefficients associated with bill length, bill depth, and bosy mass are all positive, larger penguins are more likely to be male.

(e)

A couple of interpretations:

- For a 1-mm increase in bill length, the log odds of a penguin being male increases by 0.9513, while controlling for bill depth, body mass, and species.
- For a 1-mm increase in bill length, the odds of a penguin being male is multiplied by $\exp(0.9513) = 2.589$, while controlling for bill depth, body mass, and species.

(f)

```
newdata<-data.frame(species="Gentoo", bill_length_mm=49, bill_depth_mm=15, body_mass_g=5
predict(reduced,newdata)
```

```
##        1
## 6.462668
```

```
##convert to odds
odds<-exp(predict(reduced,newdata))
odds
```

```
##        1
## 640.7683
```

```
##convert odds to probability
prob<-odds/(1+odds)
prob
```

```
##         1
## 0.9984418
```

The log odds of this penguin being male is 6.462668. The corresponding odds is 640.7683, and the corresponding probability is 0.9984418.

(g)

```
deltaG2<-reduced$null.deviance-reduced$deviance
deltaG2
```

```
## [1] 298.4472
```

```
1-pchisq(deltaG2,5)
```

```
## [1] 0
```

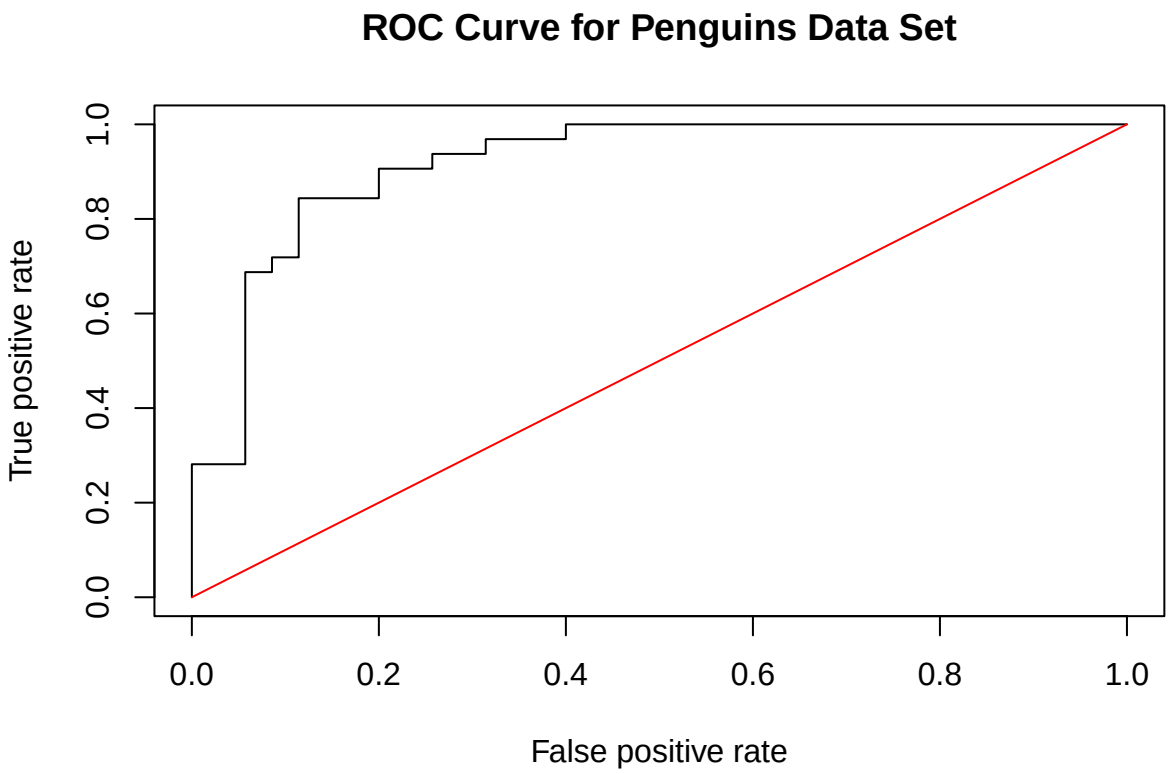
$H_0 : \beta_1 = \cdots = \beta_5 = 0$

H_a : at least one of the coefficients in H_0 is not 0.

The test statistic is 298.4472 with a p-value that is virtually 0. So we reject the null hypothesis, our model is useful.

(h)

```
library(ROCR)
##generate ROC curve
preds<-predict(result,newdata=test, type="response")
rates<-prediction(preds, test$sex)
roc_result<-performance(rates,measure="tpr", x.measure="fpr")
plot(roc_result, main="ROC Curve for Penguins Data Set")
lines(x = c(0,1), y = c(0,1), col="red")
```



Since our ROC curve is closer to the top left than the red diagonal line, our logistic regression performs better than random guessing on the test data.

(i)

```
auc<-performance(rates, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.9169643
```

Since the AUC is above 0.5, our logistic regression performs better than random guessing on the test data.

(j)

```
table(test$sex, preds>0.5)
```

```
##
##          FALSE TRUE
##   female    28    7
##   male       4   28
```

- The FPR is $\frac{7}{35} = 0.2$.
- The FNR is $\frac{4}{32} = 0.125$.
- The error rate is $\frac{7+4}{35+32} = 0.164$.

(k)

For this particular analysis, I do not believe there is a reason to favor reducing FPR over FNR (or vice versa); there isn't a worse consequence of wrongly identifying a female penguin as male versus wrongly identifying a male penguin as female. Thus we would prefer to reduce the overall error rate, which is achieved with a threshold of 0.5. So we do not need to adjust the threshold.