Problem set 11

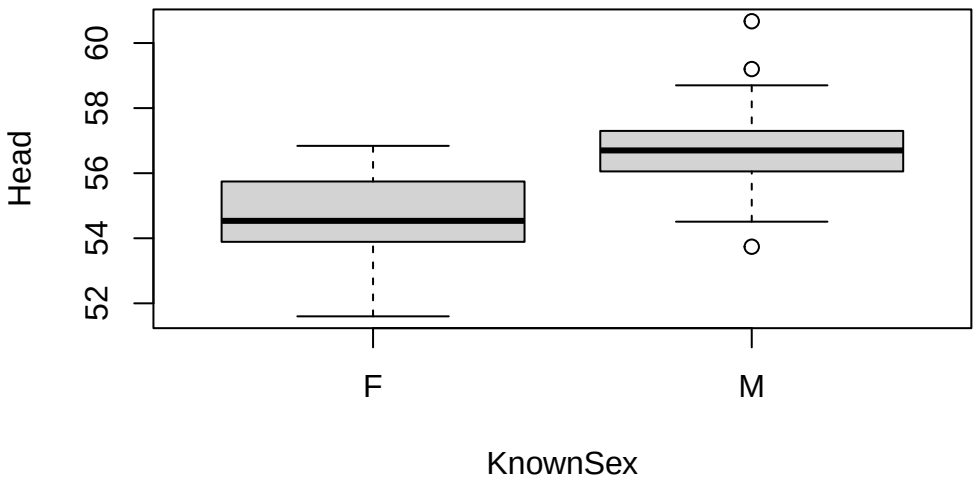
Exercises to hand in: 10.20, 10.23, 10.24

10.20 Blue jay morphology

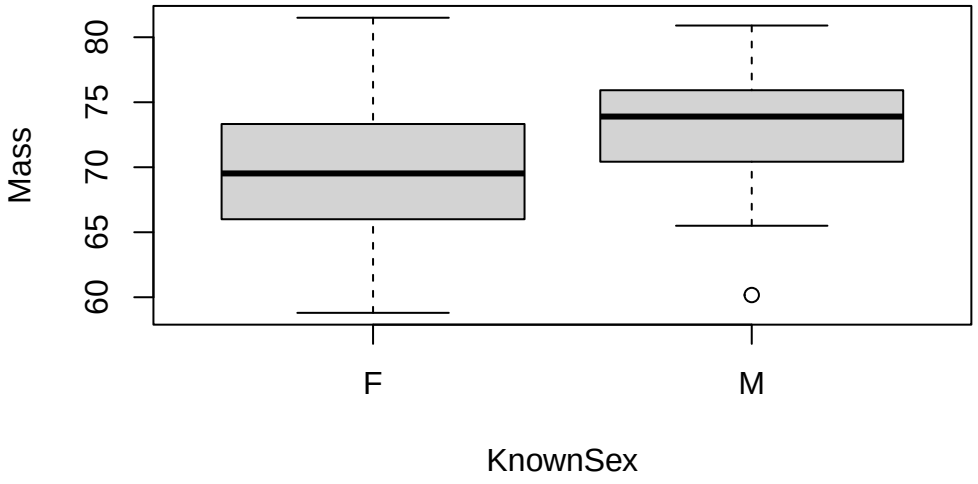
```
data("BlueJays")
```

a. Boxplots

```
attach(BlueJays)
boxplot(Head~KnownSex)
```




```
boxplot(Mass~KnownSex)
```



b. Simple logistic regression

```
log2 <- glm(Sex ~ Mass, data = BlueJays, family = binomial)
summary(log2)
```

```
Call:
glm(formula = Sex ~ Mass, family = binomial, data = BlueJays)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9514  -1.0402   0.6611   0.9985   2.0192

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.17171    3.24990  -3.745 0.000180 ***
Mass          0.17073    0.04528   3.770 0.000163 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 170.44  on 122  degrees of freedom
Residual deviance: 153.56  on 121  degrees of freedom
AIC: 157.56

Number of Fisher Scoring iterations: 4
```

```
exp(coef(log2))
```

```
 (Intercept)         Mass
5.174825e-06 1.186168e+00
```

he p-value of mass is 0.000163 which is less than 0.05. This means that mass is a significant predictor in determining the sex of blue jay. This means that mass has predicting power. If the p-value would have been greater than 0.05, then mass would have no predictive power.

c. Multiple logistic regression

```
log3<-glm(Sex ~ Mass+Head, data = BlueJays, family = binomial)
summary(log3)
```

```
Call:
glm(formula = Sex ~ Mass + Head, family = binomial, data = BlueJays)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3174  -0.5250   0.1034   0.6400   2.2532

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -89.04780   15.87609  -5.609 2.04e-08 ***
Mass         -0.07609    0.06815  -1.116    0.264
Head          1.69650    0.32271   5.257 1.46e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 170.441  on 122  degrees of freedom
Residual deviance:  99.965  on 120  degrees of freedom
AIC: 105.96

Number of Fisher Scoring iterations: 5
```

```
exp(coef(log3))
```

```
 (Intercept)         Mass         Head
2.123410e-39 9.267321e-01 5.454831e+00
```

The p-value of mass is 0.264 which is greater than 0.05 so this means that mass is not a significant predictor in determining the sex of blue jay when we do a multiple regression involving both mass and head.

10.23 Sinking of the Titanic

```
data("Titanic")
```

a. Multiple logistic

```
logisticTitanic1 <- glm(Survived ~ Age+SexCode, data=Titanic, family=binomial)
summary(logisticTitanic1)
```

```
Call:
glm(formula = Survived ~ Age + SexCode, family = binomial, data = Titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7541  -0.6905  -0.6504   0.7576   1.8628

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.159839   0.219651  -5.280 1.29e-07 ***
Age         -0.006352   0.006187  -1.027    0.305
SexCode      2.465996   0.178455  13.819  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1025.57  on 755  degrees of freedom
Residual deviance:  795.59  on 753  degrees of freedom
  (557 observations deleted due to missingness)
AIC: 801.59

Number of Fisher Scoring iterations: 4
```

$$logit(\hat{\pi}) = -1.160 - 0.006352 * Age + 2.465996 * SexCode$$

$$\hat{\pi} = \frac{e^{-1.160-0.006352*Age+2.465996*SexCode}}{1+e^{-1.160-0.006352*Age+2.465996*SexCode}}$$

b. Predictor effectiveness

Since the p-value of sex is less than 0.05, it is a significant predictor in predicting the response variable Survived whereas the p value of Age is greater than 0.05 so it is not a significant predictor in predicting the response variable Survived.

c. Predictions for 18-year-old man

```
exp(-1.160-0.00635*(18))
```

```
[1] 0.2796266
```

```
0.2796266/1.279627
```

```
[1] 0.218522
```

$\widehat{odds} = 0.2796266$ $\hat{\pi} = 0.218522$

d. Predictions for 18-year-old woman

```
exp(1.306157-0.00635*(18))
```

```
[1] 3.293191
```

```
3.293191/4.293191
```

```
[1] 0.767073
```

$\widehat{odds} = 3.293191$

$\hat{\pi} = 0.767073$

e. Redo

```
exp(-1.160-0.00635*(50))
```

```
[1] 0.2282075
```

```
exp(-1.160-0.00635*(50))/(1+exp(-1.160-0.00635*(50)))
```

```
[1] 0.1858053
```

$\widehat{odds} = 0.2282075$ $hat\pi = 0.1858053$

```
exp(1.306157-0.00635*(50))
```

```
[1] 2.687623
```

```
exp(1.306157-0.00635*(50))/(1+exp(1.306157-0.00635*(50)))
```

```
[1] 0.7288226
```

$\widehat{odds} = 2.687623$ $\hat{\pi} = 0.7288226$

```
2.687623/0.2282075
```

```
[1] 11.7771
```

The odds ratio of a woman compared to a man of the same age is 11.7771

f. Odds ratio

The gender odd ratio is the same at all ages. This will always be the case.

10.24 Titanic interaction

a. Explain coefficients

```
Titanic <- Titanic %>%
    drop_na(Age)
```

```
logisticTitanic2 <- glm(Survived ~ Age+SexCode+Age*SexCode, data=Titanic, family=binomial)
summary(logisticTitanic2)
```

```
Call:
glm(formula = Survived ~ Age + SexCode + Age * SexCode, family = binomial,
    data = Titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1262  -0.7348  -0.5194   0.7699   2.2632

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.298750   0.277699  -1.076    0.282
Age         -0.036367   0.009263  -3.926 8.63e-05 ***
SexCode      0.599858   0.408050   1.470    0.142
Age:SexCode  0.065718   0.013686   4.802 1.57e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1025.57  on 755  degrees of freedom
Residual deviance:  770.56  on 752  degrees of freedom
AIC: 778.56

Number of Fisher Scoring iterations: 4
```

```
-0.298750+0.599858
```

```
[1] 0.301108
```

```
-0.036367+0.065718
```

```
[1] 0.029351
```

log(odds_survived_female) = -0.298750 -0.036367 * Age

log(odds_survived_male) = 0.301108 + 0.029351 * Age

Since the p value is less than 0.05, this difference is statistically significant.

b. Nest likelihood ratio test

```
logisticTitanic3 <- glm(Survived ~ SexCode, data=Titanic, family=binomial)
summary(logisticTitanic3)
```

```
Call:
glm(formula = Survived ~ SexCode, family = binomial, data = Titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6735  -0.6776  -0.6776   0.7524   1.7800

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3545     0.1145  -11.83   <2e-16 ***
SexCode       2.4718     0.1783   13.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1025.57  on 755  degrees of freedom
Residual deviance:  796.64  on 754  degrees of freedom
AIC: 800.64

Number of Fisher Scoring iterations: 4
```

```
lrtest(logisticTitanic3, logisticTitanic2)
```

```
Likelihood ratio test

Model 1: Survived ~ SexCode
Model 2: Survived ~ Age + SexCode + Age * SexCode
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -398.32
2   4 -385.28  2 26.088  2.163e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p value is significant, the interaction model is better than using just the SexCode model.