M11HW

Question 1 set up

```
library(palmerpenguins)
```

```
## Warning: package 'palmerpenguins' was built under R version 4.4.3
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(gridExtra)
```

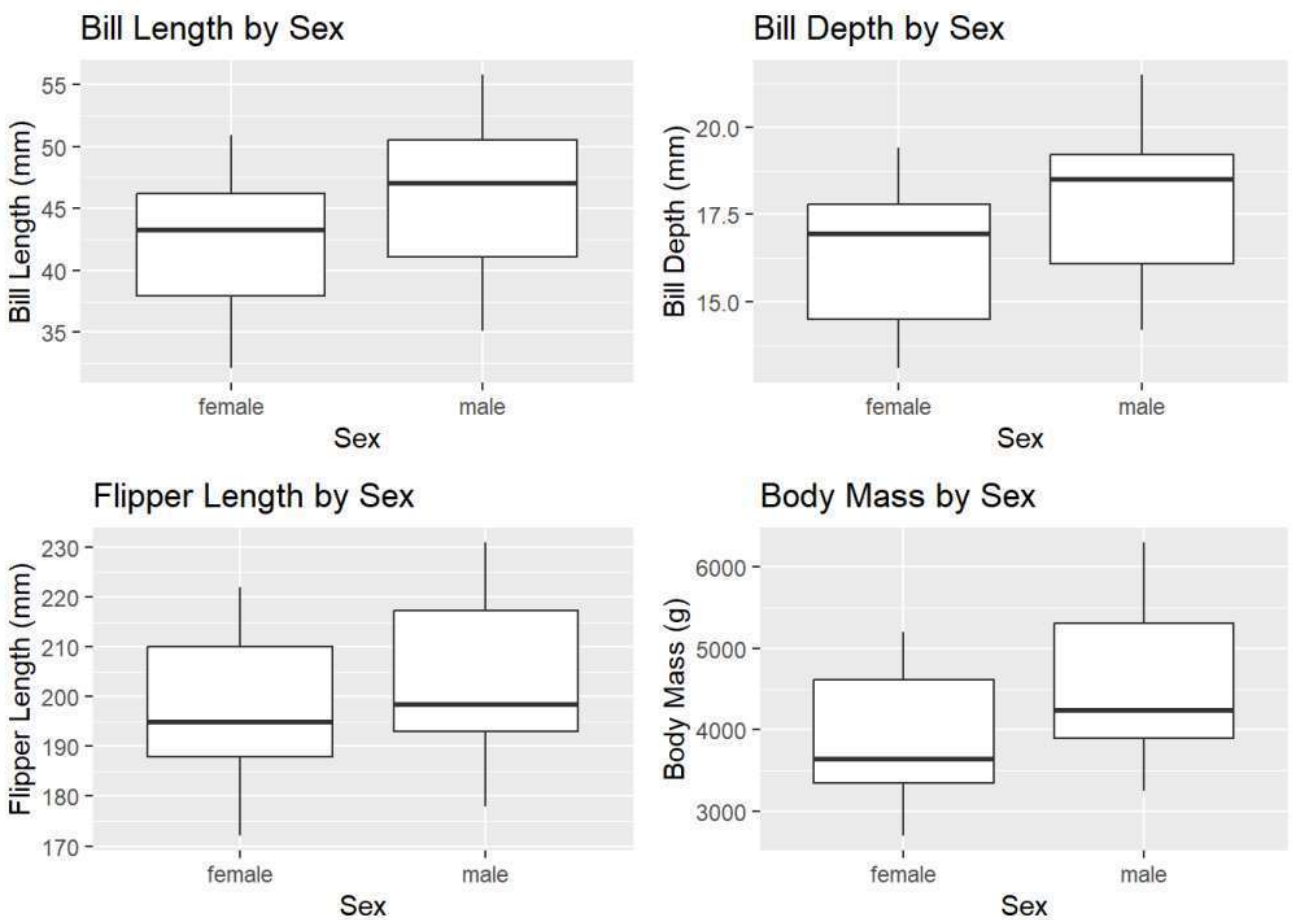
```
## Warning: package 'gridExtra' was built under R version 4.4.3
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
options(scipen=999)

Data<- penguins
## removing penguins with no gender specified and columns 2 through 8
Data<-Data[complete.cases(Data[,7]),-c(2,8)]
##80-20 split
set.seed(1) ##Always set a seed for reproduceable results
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample,] ##training data frame
test<-Data[-sample,] ##test data frame
```

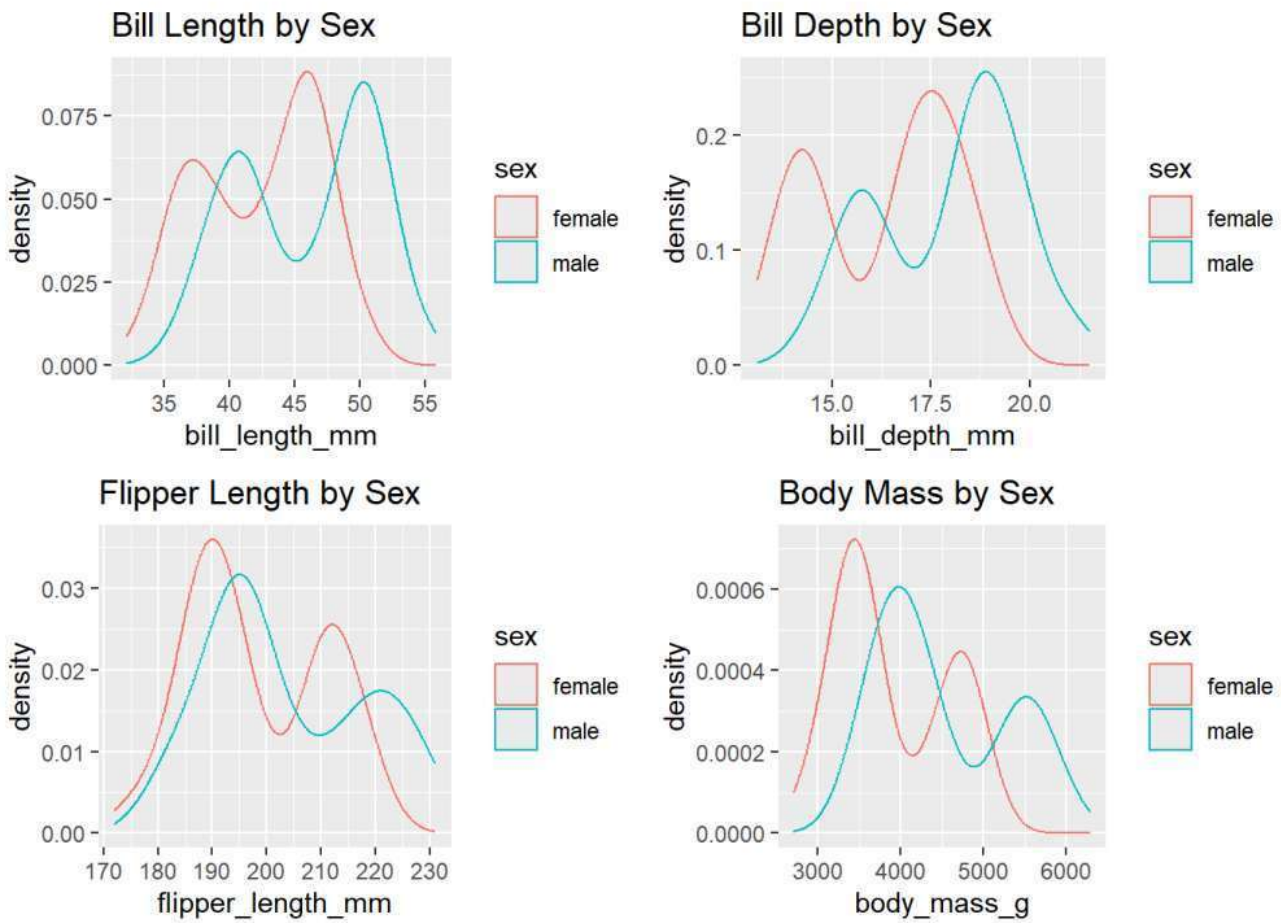
1(a) visualizations exploring relationship of measurements and gender

```
##Setting up BoxPlots by giving them a variable name, allows them to be callable and allows them
to be placed into a grid

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

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

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

bp4<-ggplot(train, aes(x=sex, y=body_mass_g))+
  geom_boxplot()+
  labs(x="Sex", y="Body Mass (g)", title= "Body Mass by Sex")

## function allows for the 4 box plots to be put in a 2 by 2 matrix
grid.arrange(bp1,bp2,bp3,bp4, ncol = 2, nrow =2)
```



Based on these box plots, it become rather evident that with at least relative confidence from a visual stance, you can make an assertion that each physical measurement is increased in males, having a sex based dependency.

```
## Setting up density plots of penguin size measurements
dp1<-ggplot(train,aes(x=bill_length_mm, color=sex))+
  geom_density()+
  labs(title="Bill Length by Sex")

dp2<-ggplot(train,aes(x=bill_depth_mm, color=sex))+
  geom_density()+
  labs(title="Bill Depth by Sex")

dp3<-ggplot(train,aes(x=flipper_length_mm, color=sex))+
  geom_density()+
  labs(title="Flipper Length by Sex")

dp4<-ggplot(train,aes(x=body_mass_g, color=sex))+
  geom_density()+
  labs(title="Body Mass by Sex")

grid.arrange(dp1,dp2,dp3,dp4, ncol=2, nrow=)
```



Looking at these plots, you see marginal densities across every single display with the largest available densities coming in Bill Depth.These densitiy measurements are definitely impacted by the changes in range of each measurements metric with some being by the thousands while others are in chunks of 5 units at a time.

However, we do see consistent patterns where there appears to more higher presence at certain measurement ranges for female penguins and certain measurement ranges that are more common for male penguins.

1(b-c) Fitting a logistic model and using the Wald test for coefficients

```
result<- glm(sex~., family = "binomial", data = train)
## Including Species to be able to control for species in the model
summary(result)
```

```
##
## Call:
## glm(formula = sex ~ ., family = "binomial", data = train)
##
## Coefficients:
##                     Estimate Std. Error z value     Pr(>|z|)
## (Intercept)       -94.355394  17.638204  -5.349 0.0000000882 ***
## speciesChinstrap  -10.608813   2.634752  -4.026 0.0000566148 ***
## speciesGentoo     -10.384568   3.565641  -2.912      0.00359 **
## bill_length_mm      1.025200   0.238593   4.297 0.0000173241 ***
## bill_depth_mm       2.287977   0.516595   4.429 0.0000094688 ***
## flipper_length_mm  -0.088318   0.065040  -1.358      0.17450
## body_mass_g         0.008094   0.001662   4.871 0.0000011111 ***
## ---
## 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
```

Wald Test: Z = estimate/standard error(estimate)

flipper: Z = -0.088318/0.65040, = -0.1357, large P-value, remove this metric Accepting the hypothesis the estimate = 0/ could = 0 so removing from the model

Cannot determine further changes until a refit is completed.

```
result2<- glm(sex~.-flipper_length_mm, family = "binomial", data = train)
## Removed flipper length from model and refitting to evaluate further variables
summary(result2)
```

```
##
## Call:
## glm(formula = sex ~ . - flipper_length_mm, family = "binomial",
##     data = train)
##
## Coefficients:
##                     Estimate  Std. Error z value      Pr(>|z|)
## (Intercept)      -103.223133   17.058939  -6.051 0.00000000144 ***
## speciesChinstrap  -10.421609    2.544372  -4.096 0.00004204479 ***
## speciesGentoo     -12.384034    3.382911  -3.661      0.000251 ***
## bill_length_mm      0.951263    0.221050   4.303 0.00001682064 ***
## bill_depth_mm       2.099138    0.468401   4.481 0.00000741208 ***
## body_mass_g         0.007714    0.001625   4.746 0.00000207069 ***
## ---
## 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:  70.172  on 260  degrees of freedom
## AIC: 82.172
##
## Number of Fisher Scoring iterations: 8
```

This is a reasonable model to not run a walds test on at the moment as the p-values for each estimate are much lower than 0.05.

Logistic Regression Equation:

log(pihat/(1-pihat)) = -103.223 + 0.9513*bill_length* + 2.0991*bill_depth* + 0.0077*body_mass* - 10.4216*I1* -12.3849*I2*

where: I1 is 1 for Chinstrap penguins and 0 for all other species and I2 is 1 for Gentoo Penguins and 0 for all other species.

1(d) how do you interpret the regression equation

Generally, as these predictors increase in their respective units, the log odds of the penguin being a male increases by the amount described by their coefficient when controlling for species and other variables.

1(e) interpret esitmate for bill length

For every millimeter of bill length, the log odds of that penguin being male increases, when controlling for all other variables and species, by 0.9513, while the estimated odds are increased by a factor of 2.589073 times for each additional mm.

Accidentally interpreted them all before I red the question

Based on this equation, it appears that bill_depth is going to be the strongest indicator for if a penguin, when controlling for species and other variables, is male or female, with higher values for this predictor making it more likely to be male, something we can extrapolate using the regression combined with the visualizations in which males tended to be larger. Meaning that the log odds of this value increases by 2.0991 for each additional mm bill depth increase when controlling for all other variables.

This pattern can also be followed with bill length and body mass with the log odds increasing by 0.9513 and 0.0077 per unit increase respectively.

In other words, the estimated odds of a penguin being male increases by a factor of 8.1588, 2.5891, and 1.077 with respect to the order they are interpreted in this description when controlling for the other variables for each one.

1(f) Make a prediction about the log-odds, odds, and probability of a Gentoo penguin's sex

```
levels(Data$species)
```

```
## [1] "Adelie"    "Chinstrap" "Gentoo"
```

```
new_gentoo<-data.frame(bill_depth_mm=15, bill_length_mm=49, body_mass_g=5700, flipper_length_mm=
220, species="Gentoo")

##making prediction for log odds
log_odds<-predict(result2,new_gentoo)
log_odds
```

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

```
## prediction for odds
odds<- exp(log_odds)
odds
```

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

```
prob<- odds/(1+odds)
prob
```

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

Based on the outcomes from the model, the log odds are 6.46… times more likely to be a male, with the odds being 640.7683 to 1 that this is a male, meaing it is most likely a male, at a probability of 99.84…% that this penguin is a Gentoo male penguin.

1(f) conducting a hypothesis test to assess how useful the logistic regression is

```
##likelihood ratio test
deltaG2<- result2$null.deviance-result2$deviance
deltaG2
```

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

```
1-pchisq(deltaG2, 5) # df = 5 including species to go from full model to null reduced model, uns
ure if this value should actually be, but same answer both ways.
```

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

Null: at least one of the coefficients = 0 Alternative: none of the coefficients = 0

with a test statistic 298.4472, the p-value is 0, so we reject the null and assert that the 4 predictor model is significantly more useful than the intercept model and assert that at least one of the coefficients is a nonzero.

HW11 Question 2(a)

The coefficient for X3 can be interpreted that a male, who's encoded as a 1, will have log-odds 0.43397 increase by 0.43397, with their flat odds being 1.548373 times higher of having gotten the flu shot than that of an elderly female.

HW11 2(b)

$H_0: \beta_3 = 0$ $H_a: \beta_3 \neq 0$

$0.43397 / 0.52179 = 0.8316947 = Z$

$2*(1-pnorm(0.8316947))$

p-value $\approx 0.40056 > 0.05$

Accept the null hypothesis allowing us to drop gender from the model.

HW11 2(c)

$\hat{\beta}_3 \pm Z_{1-\alpha/2} Se(\hat{\beta}_3)$

$0.43397 \pm (0.8316947)(0.52179)$

$(-0.000...022487, 0.86794)$

~~While it is technically in range to not~~

This means the ~~Coefficient for~~ β_3 will fall somewhere between a very near Zero value an 0.868 95% of the time.

HW 11 2(d)
At the most strict interpretation, these conclusions from these tests are contradictory. However, the lower bound is so close to zero, that it could easily be interpreted as if it were in fact 0 following the results of the hypothesis test.

HW 11 2(e) $H_0: \beta_2 = 0$ $H_a: \beta_2 \neq 0$

$\Delta G2 = 134.94 - 113.20 = 21.74$

$1 - pchisq(21.74, 2) = 1.9 \times 10^{-5}$

$1.9 \times 10^{-5} < 0.05$, Reject the null and assert this as a sufficient model for predicting flu shot status supporting the reduced model.

HW 11 2(f)

$\log\left(\frac{p}{1-p}\right) = 4.911338 - 0.11931(65)$

log-odds $= -2.84382$

odds $= \exp(-2.84382)$

Odds $= 0.05820291$

probability $= \frac{0.05820291}{1-0.5820291}$

$\approx 0.0618 \approx 6.18\%$ chance that this client got the flu shot.

Although this is what the model predicts, it does not make sense that individuals w/ higher health awareness score would be predicted to be less likely to have had their flu shot suggesting errors in data description and/or collection.