uwa6xv_M11_HW

Alanna Hazlett

2024-04-17

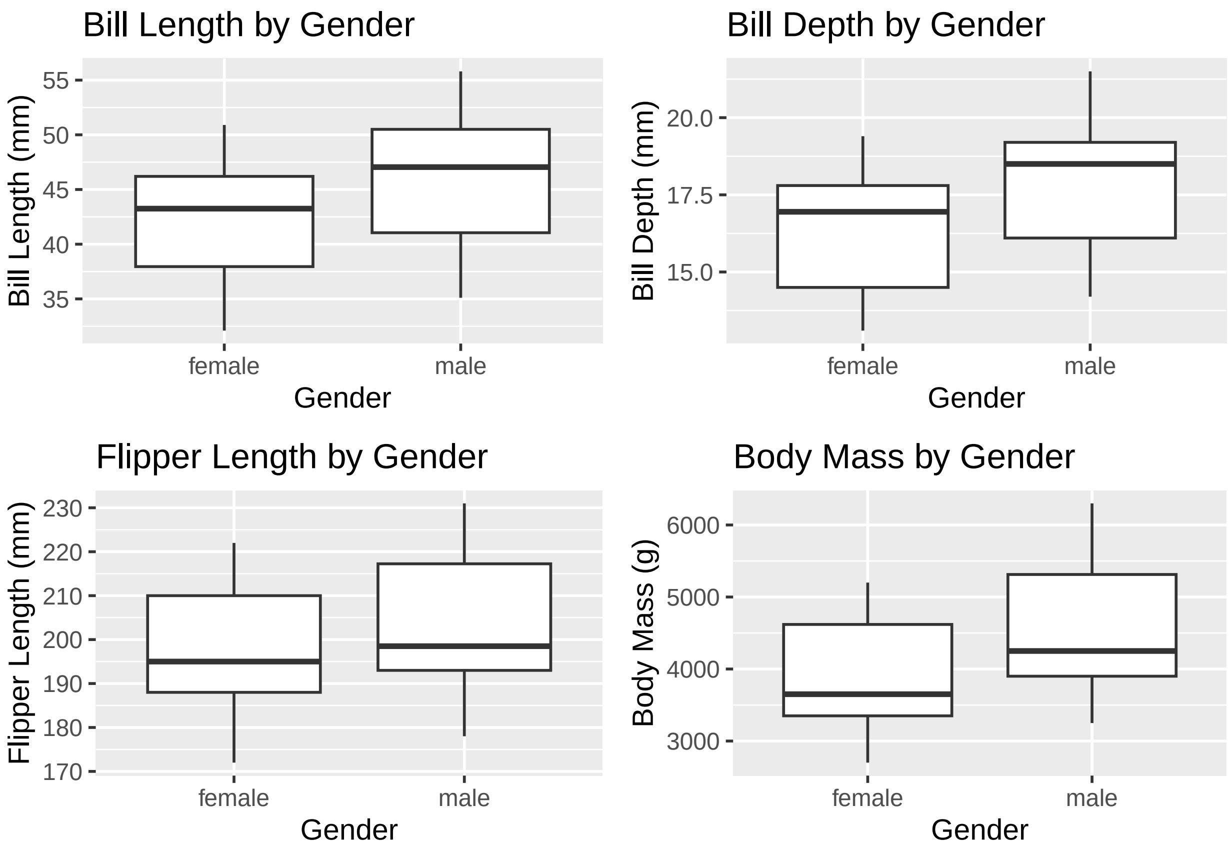
Problem 1

```r
library(palmerpenguins)
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, ]
```

(a)

Create some data visualizations to explore the relationship between the various body measurements and the gender of penguins. Be sure to briefly comment on your data visualizations.

```r
bp1<-ggplot(train, aes(x=sex, y=bill_length_mm))+
  geom_boxplot()+
  labs(x="Gender", y="Bill Length (mm)", title="Bill Length by Gender")
bp2<-ggplot(train, aes(x=sex, y=bill_depth_mm))+
  geom_boxplot()+
  labs(x="Gender", y="Bill Depth (mm)", title="Bill Depth by Gender")
bp3<-ggplot(train, aes(x=sex, y=flipper_length_mm))+
  geom_boxplot()+
  labs(x="Gender", y="Flipper Length (mm)", title="Flipper Length by Gender")
bp4<-ggplot(train, aes(x=sex, y=body_mass_g))+
  geom_boxplot()+
  labs(x="Gender", y="Body Mass (g)", title="Body Mass by Gender")
grid.arrange(bp1, bp2, bp3, bp4, ncol = 2, nrow = 2)
```


Bill Length by Gender

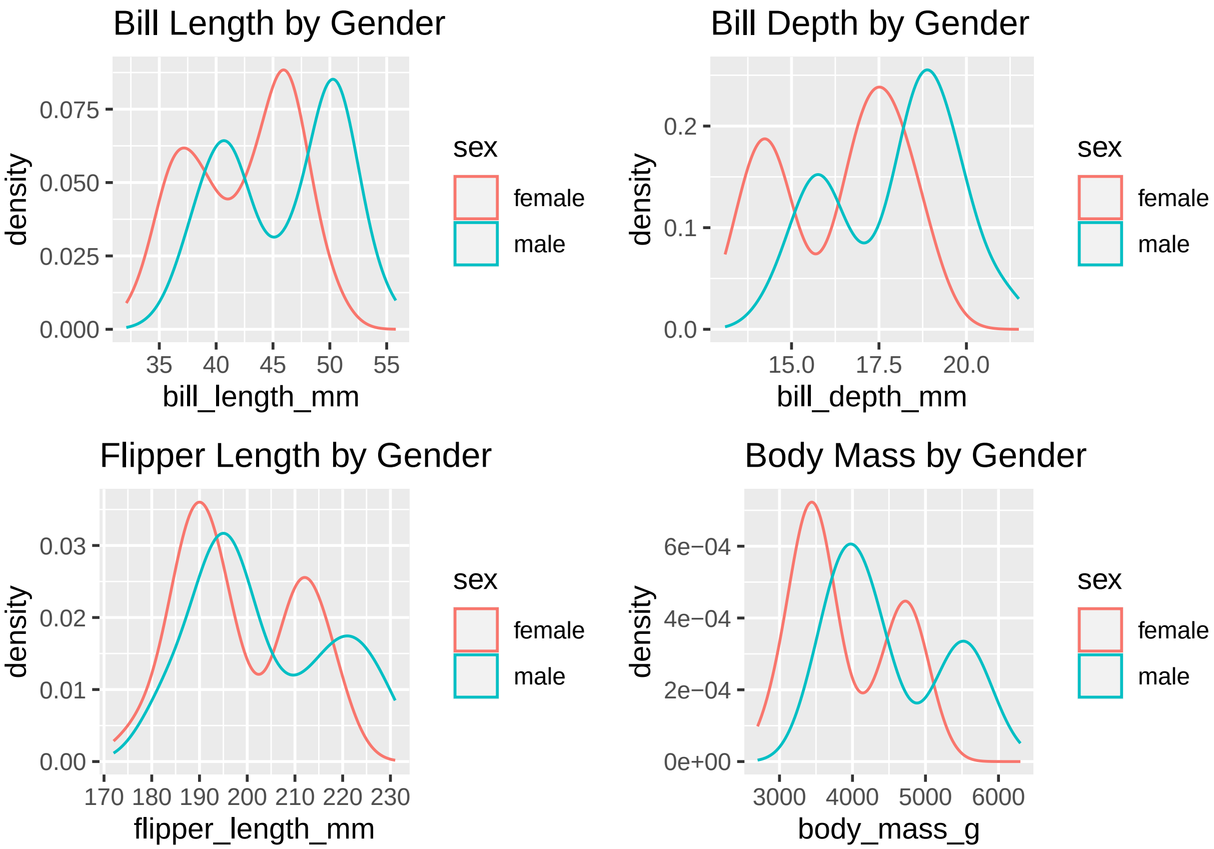
Bill Depth by Gender

Flipper Length by Gender

Body Mass by Gender

For all body measurements we are investigating, the males have higher values than the females. The variance of the values is fairly similar comparing females and males for each of the measurements.

```
dp1<-ggplot2::ggplot(train,aes(x=bill_length_mm, color=sex))+
  geom_density()+
  labs(title="Bill Length by Gender")
dp2<-ggplot2::ggplot(train,aes(x=bill_depth_mm, color=sex))+
  geom_density()+
  labs(title="Bill Depth by Gender")
dp3<-ggplot2::ggplot(train,aes(x=flipper_length_mm, color=sex))+
  geom_density()+
  labs(title="Flipper Length by Gender")
dp4<-ggplot2::ggplot(train,aes(x=body_mass_g, color=sex))+
  geom_density()+
  labs(title="Body Mass by Gender")
gridExtra::grid.arrange(dp1, dp2, dp3, dp4, ncol = 2, nrow = 2)
```



We can see in all four measurements that the males have higher values and the distribution of these values are rather similar from female to male, except for bill depth. We can see this based on the shape of the density line, they are very similar. In bill depth there are two peaks for females but they are not as different in density as the males.

(b)

Use R to fit the logistic regression model. Based on the results of the Wald tests for the individual coefficients, which predictor(s) appears to be insignificant in the model?

```
result<-glm(sex~bill_length_mm+bill_depth_mm+flipper_length_mm+body_mass_g,family=binomial,data=train)
summary(result)
```

```
##
## Call:
## glm(formula = sex ~ bill_length_mm + bill_depth_mm + flipper_length_mm +
##     body_mass_g, family = binomial, data = train)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -60.715378  10.487887  -5.789 7.08e-09 ***
## bill_length_mm     0.151168   0.063371   2.385   0.0171 *
## bill_depth_mm      2.460582   0.349778   7.035 2.00e-12 ***
## flipper_length_mm -0.086560   0.043632  -1.984   0.0473 *
## body_mass_g        0.007025   0.001153   6.093 1.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 368.62  on 265  degrees of freedom
## Residual deviance: 103.93  on 261  degrees of freedom
## AIC: 113.93
##
## Number of Fisher Scoring iterations: 7
```

The p-values for bill length, bill depth, and body mass are small, which means for each Wald test for each coefficient we would reject the null hypothesis of the coefficient equaling zero, our data supports the alternative that each coefficient does not equal zero, meaning that it is useful for our model, in the presence of the other predictors.For flipper length the p-value is in a gray area, but when we look at the fact that the estimate coefficient is negative and the standard error isn't large enough to change it to be a positive slope we can determine that it is not beneficial for our model. This is because from our previous visualizations we see that increasing flipper length does show higher probability of the penguin being male, indicating that the slope of the coefficient should be positive.

(c)

Based on your answer in part 1b, drop the predictor(s) and refit the logistic regression. Write out the estimated logistic regression equation. If you did not drop any predictor, write out the logistic regression equation from part 1b.

```r
result.reduced<-glm(sex~bill_length_mm+bill_depth_mm+body_mass_g,family=binomial,data=train)
summary(result.reduced)
```

```
##
## Call:
## glm(formula = sex ~ bill_length_mm + bill_depth_mm + body_mass_g,
##     family = binomial, data = train)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.064e+01  9.543e+00  -7.403 1.34e-13 ***
## bill_length_mm  9.558e-02  5.501e-02   1.737   0.0823 .
## bill_depth_mm   2.482e+00  3.380e-01   7.342 2.11e-13 ***
## body_mass_g     5.756e-03  8.397e-04   6.855 7.11e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 368.62  on 265  degrees of freedom
## Residual deviance: 108.04  on 262  degrees of freedom
## AIC: 116.04
##
## Number of Fisher Scoring iterations: 7
```

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -70.6428079 + 0.0955796x_{BillLength} + 2.4816625x_{BillDepth} + 0.0057562x_{BodyMass}$$

(d)

Based on your estimated logistic regression equation in part 1c, how would you generalize the relationship between some of the body measurement predictors and the (log) odds of a penguin being male?

```r
contrasts(train$sex)
```

```
##        male
## female    0
## male      1
```

Log odds increases for the penguin being male when bill length, bill depth, or body mass increase.

(e)

Interpret the estimated coefficient for bill length contextually.
Estimated coefficient for bill length is 0.0955796.
Estimated log(odds) of the penguin being male increases by 0.0955796 for each additional mm increase in bill length (on average), when controlling for the other predictors.
Estimated odds for the penguin being male is multiplied by e^0.0955796 = 1.1003 for each additional mm increase in bill length (on average), when controlling the other predictors.

(f)

Consider a Gentoo penguin with bill length of 49mm, bill depth of 15mm, flipper length of 220mm, and body mass of 5700g. What are the log odds, odds, and probability that this penguin is male?

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -70.6428079 + 0.0955796(49) + 2.4816625(15) + 0.0057562(5700)$$

```r
##make prediction for log odds
newdata<-data.frame(bill_length_mm=49, bill_depth_mm=15, flipper_length_mm=220, body_mass_g=5700)
predict(result.reduced,newdata)
```

```
##        1
## 4.076052
```

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

```
##        1
## 58.91242
```

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

```
##        1
## 0.983309
```

(g)

Conduct a relevant hypothesis test to assess if the logistic regression in part 1c is a useful model. Be sure to write out the null and alternative hypotheses, report the value of the test statistic, and write a relevant conclusion.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_a : at\ least\ one\ coefficient\ \neq 0$$

Test statistic Delta G^2:

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

```
## [1] 260.5779
```

```r
#p-value for chi squared dist.
1-pchisq(deltaG2,4)
```

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

P-value is beyond small and is zero, we reject the null hypothesis that our coefficients are equal to zero. Our data supports the alternative hypothesis that at least one of the coefficients is not equal to zero and supports that our model is useful.

Problem 2

(a)

Interpret the estimated coefficient for x3, gender, in context.
Where males were coded x3 = 1 and females were coded x3 = 0 Coefficient = 0.43397.
This indicates the difference from females to males. The estimated log(odds) is 0.43397 higher for males than females, when controlling the other predictors.
The estimated odds for males is e^0.43397 = 1.5434 times higher than the odds for females, when controlling for the other predictors.

(b)

Conduct the Wald test for Beta3. State the null and alternative hypotheses, calculate the test statistic, and make a conclusion in context.
$H_0 : \beta_3 = 0$ $H_a : \beta_3 \neq 0$
Z statistic:

$$Z = \frac{\hat{\beta}_j - value\ in\ null\ hypothesis}{se(\hat{\beta}_j)} = \frac{0.43397 - 0}{0.52179} = 0.8317$$

```
#p-value
2*(1-pnorm(abs(0.8317)))
```

```
## [1] 0.4055783
```

Our p-value is large, we fail to reject the null hypothesis that beta3 = 0. We should drop x3 (gender) while keeping the other predictors in the model.

(c)

Calculate a 95% confidence interval for Beta3, and interpret the interval in context.

$$CI = \hat{\beta}_j \pm Z_{1-(\frac{\alpha}{2})} * se(\hat{\beta}_j)$$

```
#Get Z multiplier
qnorm(1-(0.05/2))
```

```
## [1] 1.959964
```

$$95\% CI = 0.43397 \pm (1.959964 * 0.52179) = \ (-0.5887, 1.4566)$$

Zero is within our interval, so we should drop this coefficient from our model.

(d)

Comment on whether your conclusions from parts 2b and 2c are consistent.
My conclusions are consistent. This what we would expect as both are ways to determine if a coefficient is effective in the model.

(e)

Suppose you want to drop the coefficients for age and gender, Beta1 and Beta3. A logistic regression model for just awareness was fitted, and the output is shown below.

$$H_0 : \beta_1 = \beta_3 = 0$$

$$H_a : at\ least\ one\ coefficient\ \neq 0$$

$$\Delta G^2 = D(R) - D(F) = 113.20 - 105.09 = 8.11$$

```
1-pchisq(8.11,2)
```

```
## [1] 0.01733548
```

Our p-value is small, we reject the null hypothesis. The data supports the alternative hypothesis that at least one of these coefficients does not equal zero and that they are useful for our model.

(f)

Based on your conclusion in question 2e, what are the estimated odds of a client receiving the flu shot if the client is 70 years old, has a health awareness rating of 65, and is male? What is the estimated probability of this client receiving the flu shot?

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -1.17716 + 0.07279x_{age} + -0.09899x_{aware} + 0.43397x_{gender}$$

The log(odds):

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -1.17716 + 0.07279(70) + -0.09899(65) + 0.43397(1) = \ -2.08224$$

The estimated odds:

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{-1.17716+0.07279(70)+-0.09899(65)+0.43397(1)= \ -2.08224} = 0.1247$$

The estimated probability:

$$\hat{\pi} = \frac{e^{-1.17716+0.07279(70)+-0.09899(65)+0.43397(1)}}{1+e^{-1.17716+0.07279(70)+-0.09899(65)+0.43397(1)}} = \frac{0.1246506817}{1+0.1246506817} = 0.1108$$